Predicting Bird Occurrences from High-resolution Aerial Images

Laurel Hopkins
Oregon State University
Corvallis, Oregon, USA
hopkilau@oregonstate.edu

Ulises Zaragoza
Oregon State University
Corvallis, Oregon, USA
zaragozu@oregonstate.edu

Rebecca Hutchinson
Oregon State University
Corvallis, Oregon, USA
rah@oregonstate.edu

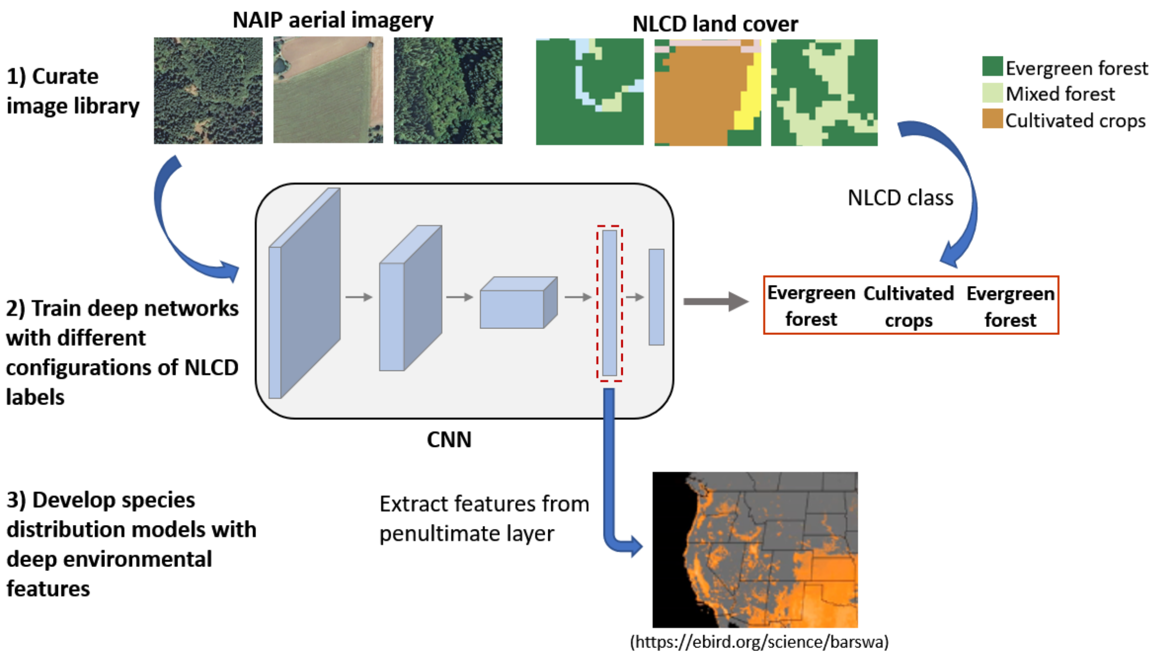

Figure 1: Method for extracting deep environmental features from high-resolution aerial images with with convolutional neural network

ABSTRACT

Species distribution models (SDMs) link species observations to environmental variables and are crucial tools in biodiversity conservation for informing land management tasks. Current methods for collecting environmental variables generally consist of field-based methods or basic summaries of remotely sensed data. As the spatial extents of models expand to map continental and global patterns, field-based methods will not be feasible for collecting sufficient data. Given the advancements in computer vision due to deep learning, features extracted by deep networks have the potential to characterize habitats better than methods currently used in SDM. We implemented a multistage transfer learning approach to map species occurrences to aerial images. We evaluated the performance of the deep environmental feature sets by predicting five bird distributions in Oregon with records from the citizen science-based eBird dataset. Surprisingly, we find the deep environmental features are no more informative of species occurrences than simple land cover summaries and explore several possible reasons for these unexpected results.

KEYWORDS

Machine Learning, Deep Learning, Convolutional Neural Networks, Transfer Learning, Remote Sensing, High-resolution Aerial Images, Species Distribution Models, Computational Sustainability

1 INTRODUCTION

Remote sensing data provides a wealth of information for environmental applications such as modeling species distributions and predicting water quality. Remotely sensed data products collected from programs such as Landsat, MODIS, and Sentinel are not only available at high spatial and temporal resolutions, but they are free and easily accessible which make them good alternatives to costly *in situ* measurements.

In the past decade, remote sensing imagery has become an increasingly popular source for obtaining environmental features for species distribution models (SDMs) [4, 12, 18]. Buermann et al. (2008) found improved performance in SDMs across taxonomic groups using features derived from remote sensing data, with remote sensing features especially useful in homogeneous climates spaces. Shirley et al. (2013) found that SDMs based on summaries of raw multispectral imagery are highly predictive of species occurrences with bands associated with vegetation biomass and photosynthetic activity being the most informative. To date, remotely sensed data used in SDMs are summarized with basic statistics. Buermann et al. (2008) summarized multispectral images by annual maximum, minimum, mean, standard deviation, and range, and Shirley et al. (2013) summarized similar images by mean and standard deviation.

Current methods for summarizing remote sensing imagery for SDMs are rudimentary compared to state-of-the-art computer vision techniques. Basic statistical summaries do not sufficiently capture elaborate habitat characteristics, such as habitat configuration (i.e., the same amount of habitat but arranged in different patterns). Capturing the complex spatial distributions of habitats is compulsory for the conservation of species sensitive to habitat change [1]. Understanding how the spatial configuration of habitats affects species abundances is important for land management tasks such as restoration (is it better to create a narrow corridor between habitat patches or is it better to allocate those resources to expanding patches to create less fragmented habitats?). The configuration of habitats can impact numerous phenomena, such as migration patterns, breeding, and food availability, all of which influence species distributions [2]. Habitat configuration is of special concern in Oregon as many contiguous forest patches are fragmented by clear cutting (Figure 2). Summarizing habitats with basic metrics is an oversimplification of landscapes and prevents models from determining how characteristics such as habitat configuration impacts species distributions. In contrast, our method has the potential to develop more informative habitat summaries which, in theory, will lead to more accurate species distribution models enabling better biodiversity conservation.

Figure 2: Habitat amount versus configuration; basic statistics do not capture habitat configuration.

Rather than characterizing habitats using coarse statistics of remotely sensed data, we explore SDMs developed with habitat summaries derived from deep environmental features. The versatility of deep features in downstream modeling tasks (i.e. feature learning) has been highlighted by classifying images and videos with features extracted from networks trained on relatively different sets of data (e.g., classifying human activities with features extracted from networks trained to classify sports clips) [11, 17]. Although a majority of previous work in benchmarking the transferability of deep features has focused on object-centric images, more recent work has shown similar finding with aerial images [20].

We employ a multistage transfer learning approach to extract environmental features from high-resolution aerial images to be used in the downstream task of predicting species occurrences. In the first stage, we extract features from remotely sensed images by fine-tuning a pretrained convolutional neural network (CNN) on a proxy task of predicting land cover classes from RGB aerial images. In the second stage, we build SDMs with the extracted deep features to predict species occurrences. As applying deep learning methods to remote sensing data is largely unexplored [10], we develop several configurations of deep environmental features to understand which architecture choices provide the best habitat summaries for species modeling. Specifically, we compare features extracted from a complex architecture (ResNet) [7] to features extracted from a significantly less complex architecture (AlexNet) [13]. We develop species models with records obtained from the citizen science-based eBird dataset. We benchmark our deep environmental features by modeling five species in Oregon in the year 2011 and compare our deep feature sets to complex habitat summaries provided with the eBird records (*eBird*) and a more basic summary method (*mean & SD*).

While we expected the deep features to outperform basic summaries of land cover, we surprisingly found that the deep environmental features had almost identical performances to the simpler feature sets. Additionally, and even more surprisingly, we found increasing the complexity of the transfer learning proxy task (different types of land cover classification) did not improve model performance in the downstream modeling task. We also witnessed comparable results for deep environmental features extracted with ResNet and AlexNet architectures. These unexpected findings are a reminder that increased architecture complexity is not always necessary. Below, we describe further experiments to elucidate potential factors influencing these results and suggest areas of future work to further clarify the appropriate architecture complexity for particular problem characteristics.

2 RELATED WORK

Our method is inspired by previous work on predicting poverty from satellite images. Given insufficient labels for predicting socioeconomic data in developing countries directly from satellite images, Xie et al. (2016) used NOAA's nighttime light intensities as the proxy task for fine-tuning a CNN to extract features relevant to socioeconomic status and subsequently trained a linear classifier on the extracted features to predict poverty metrics. Our task of predicting species occurrences is also limited in data set size which prevents us from predicting species occurrences directly from aerial images. Where Xie et al. (2016) used nighttime light intensity as their proxy task for poverty, we use land cover class as a proxy to determine if a species is present.

Other related work in applying deep learning methods to remote sensing datasets with limited training labels includes a dimensionality reduction technique for predicting crop yields [21] and a method for learning unsupervised representations of spatial data [10]. You et al. (2017) performed dimensionality reduction on the input aerial images by transforming RGB images into histograms of pixels to reduce the complexity of the training task. Their approach assumed that the distribution of pixel values within an image is more important than the spatial configuration of the pixels for determining average crop yield. As habitat configuration is an important factor in SDMs [2, 5], we cannot make the same amount-configuration assumption.

Jean et al. (2019) developed an unsupervised method for extracting low-dimensional representations from remote sensing imagery. They formulated the problem similarly to a natural language processing task in which images sampled geographically closer to one another should have more similar low-dimensional feature embeddings than images further apart. While habitats close in proximity are likely similar, this is not guaranteed. For our purposes, land cover type and configuration would be better indicators of similarity.

3 METHODS

We adopt a two-stage transfer learning approach to go from a pretrained CNN to predicting species occurrences through an intermediate proxy task (predicting land cover classes). In the first stage, the proxy task modifies the pretrained network to predict configurations of land cover from aerial images (ImageNet to land cover). In the second stage, we extract deep features with our tuned network from aerial images centered at bird record locations. We then model the target task (species occurrence) with the extracted deep environmental features (land cover to SDM). Our complete process is outlined in Figure 1.

3.1 Stage 1: ImageNet to Land Cover

The goal of this fine-tuning stage is to extract features most relevant to SDM from aerial images. A central question is whether deep features can represent habitat configuration better than current methods, so we perform fine-tuning on increasingly challenging proxy tasks. The first and simplest task is predicting the majority land cover class for a given image. While this task trains the network to learn features relevant to overall patterns and textures in images, it does not require the network to learn features that describe land cover configuration. Next, we fine-tuned the network on the slightly more challenging task of predicting all classes present in the image. Finally, to incorporate knowledge on the amount of classes present, we fine-tuned the network to predict the proportions of all classes present within an image. For all tasks we started with a ResNet-18 architecture pretrained on ImageNet. To understand if low-level features learned for object classification are relevant to aerial images, we trained two CNNs. In our first model we froze the weights prior to layer 2, meaning the weights were not updated during backpropagation. In our second model, all weights were updated, allowing for different low-level features to be learned. We did not see a significant difference in the accuracies of the two models but to allow for flexibility, we performed backpropagation on all layers in all future models. To compare the extracted features to a simpler architecture, we also trained an AlexNet architecture pretrained on ImageNet to predict majority land cover class. See Table 1 for a full description of feature sets.

3.2 Stage 2: Land Cover to Species Occurrence

We modeled bird occurrences with Random Forests [3] and Occupancy-Detection models [14]. Occupancy-Detection models are a common approach in statistical ecology that separate the influence of habitat features from features related to the probability of detecting a species given that it is present (e.g., time of day, time spent observing). Both the habitat and detection components of the model are essentially logistic regressions, and we fit them with a ridge regression penalty [9]. For each eBird record, we obtain an environmental feature set by evaluating the fine-tuned CNN on a 2km $\times$ 2km aerial image centered at the observation location. With the extracted feature set, we create a supplemented feature set by appending the elevation of the record and four variables associated with probability of detecting the species (month of observation, time of day, hours spent observing, number of observers).

Table 1: Environmental feature sets. All deep features are extracted from the penultimate layer

Baseline features	
mean & SD	mean and standard deviation for each image band
eBird	habitat configuration statistics derived from MODIS land cover data
Deep features	
AN_class	AlexNet trained to predict majority land cover label
RN_class	ResNet trained to predict majority land cover label
RN_ind	ResNet trained to predict binary indicator of all classes present in image
RN_prop	ResNet trained to predict proportions of all classes present in image

4 DATA

We utilize two remote sensing datasets for training the deep networks. As readily available pretrained CNN architectures are trained on three-channel (RGB) images, we chose to analyze RGB aerial **National Agriculture Imagery Program (NAIP)** images over other multispectral options (Landsat, MODIS, etc.) to maintain the direct relation to RGB input channels. NAIP images are obtained during the leaf-on agricultural growing season and have 1m spatial resolution with up to 6 meter accuracy as measured by ground control points [16]. As labels for the fine-tuning task, we use land cover maps from **The National Land Cover Database (NLCD)**. NLCD land cover classifications are based on two Landsat Thematic mapper images, a leaf-on and leaf-off pair, for the year of interest and have 30-m spatial resolution to match that of Landsat. We chose NLCD land cover labels over other land cover products for its superior resolution (30m resolution compared to 500m for MODIS land cover) [8]. All remote sensing data were processed and collected through Google Earth Engine [6], a platform for easily accessing and processing remote sensing data. We initially curated an image library of over 70,000 NAIP images but downsampled the library to obtain a more even distribution of land cover classes. The final image library consisted of roughly 50,000 NAIP images from 11 land cover classes.

For species data, we obtain bird occurrences from the global, citizen-science **eBird** dataset. Users submit checklists indicating all bird species observed and include a set of detection variables (distance traveled, number of observers, etc.). When submitting records, users indicate if they are reporting all species observed; if so, absences can be inferred. Any unusual recordings (high count, rarity, out of season) are flagged and verified by expert reviewers [19].

5 EXPERIMENTS

We modeled occurrences for five resident (non-migratory) species (American Crow, American Robin, Black-capped Chickadee, European Starling & Song Sparrow) in Oregon for the year 2011. We filtered eBird records to only include stationary count types, removing data from travelling counts, exhaustive area counts, and specialized bird watches. We extracted deep features from 2km × 2km images, as Shirley et al. (2013) found that habitat features summarized at 2km radii were informative for predicting bird occurrences.

We compared our deep environmental features to two baseline feature sets, *eBird* and *mean & SD*. The *eBird* features consist of land cover configuration summaries derived from MODIS satellite data with FRAGSTATS metrics [15]. The *eBird* features are a strong baseline for determining if the deep features capture habitat configuration as the FRAGSTATS metrics quantify some characteristics of fragmentation. The *mean & SD* features are the simplest feature set, which summarize satellite images by mean and standard deviation of each image band.

We initially built SDMs for the five species with Random Forests. We performed sensitivity analysis for the parameter *mtry*, the number of features randomly sampled as possible split candidates, and *ntree*, the number of trees to build in each forest. There was a slight increase in AUC between 1k and 5k trees and no significant difference between 5k and 10k trees. We halved and doubled the default value of *mtry* ($\sqrt{p}$ where p is the number of input features) and found no significant difference in AUC. Therefore, we ran the experiments with 5k trees and the default value of *mtry*.

Our initial results (not shown) indicated that the detection variables were highly important in predicting species occurrences. In the Random Forest models, we found that feature sets with fewer habitat variables generally outperformed feature sets with more habitat variables. At each split in the process of growing tress in Random Forest, a subset of input variables are randomly selected as split candidates and the variable that provides the highest decrease in entropy is selected. Although our feature sets vary in size (13 to 71 input variables), all feature sets have the same four detection variables. This means the proportion of the feature set that is comprised of detection variables significantly varies across feature sets. The difference in the proportion of detection variables explains why smaller feature sets had better performances in Random Forest; feature sets with higher proportions of detection variables had higher chances of a detection variable being included in the subset of potential split candidates.

It is likely the importance of detection variables in predicting species occurrence is due to the species analyzed in our study. All five species analyzed have fairly general habitat preferences and occur commonly in Oregon. As more time is spent observing, the likelihood of detecting the species increases, making the predictions highly dependent on input variables like the number of hours the observer(s) spent birding. Based on this information and the fact that the feature sets vary in size, we modified the training procedure

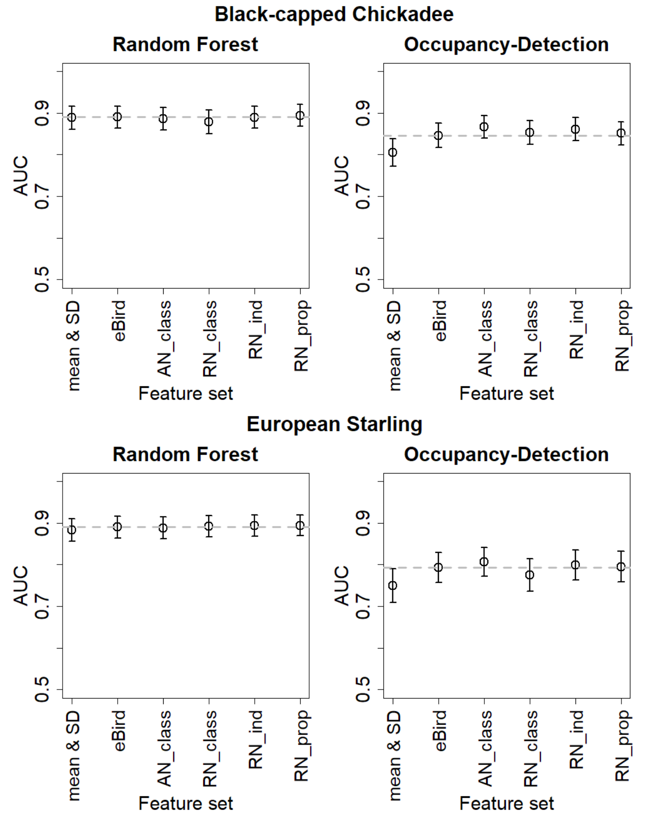

Figure 3: AUCs for each feature set modeled with Random Forest with tuned *mtry* (equal probabilities of selecting detection features as split candidates) and an Occupancy-Detection model for Black Capped Chickadee and European Starling. The dotted line represents the AUC of the baseline *eBird* feature set.

to equalize the influence of the detection variables across feature sets. Specifically, we tuned *mtry* for each feature set such that the probability of selecting a detection variable as a potential split candidate in Random Forest is equal across feature sets. Additionally, we modeled species occurrences with Occupancy-Detection models which account for detection features separately from habitat features.

6 RESULTS AND DISCUSSION

After accounting for the strong influence of detection variables, both in Random Forest and with Occupancy-Detection models, there is little difference in model performance across the habitat-based feature sets in the Random Forest models and only slight variation in the Occupancy-Detection Models (Figure 3). We only present the results for two species (Black-capped Chickadee and European Starling), as the results for the remaining species had similar trends. The most notable difference in the results is that the Random Forest models perform better than the Occupancy-Detection models. A possible explanation for the difference in performance is how the features are treated within the models; Random Forest builds

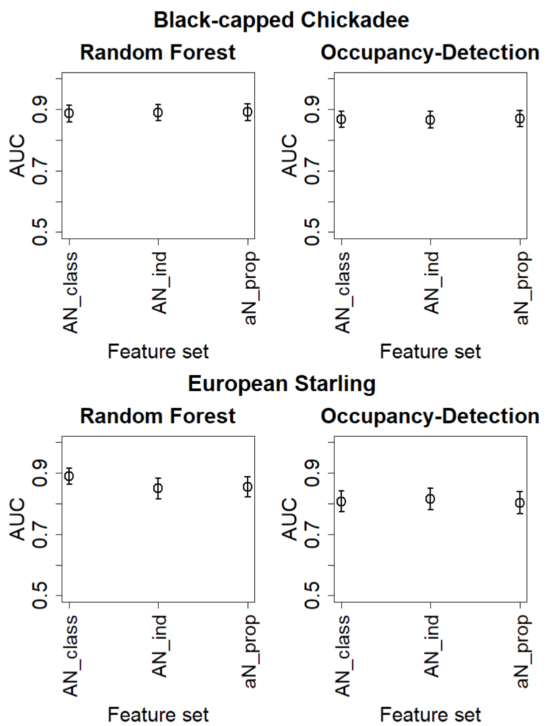

Figure 4: AUCs for the AlexNet feature sets modeled with Random Forest and an Occupancy-Detection model for Black Capped Chickadee and European Starling

a nonlinear combination of features while Occupancy-Detection models combine features linearly.

It is interesting to note that the deep feature sets extracted from AlexNet architectures perform as well or better than the ResNet-based feature sets in the Occupancy-Detection models (this trend holds across all five species). It may be the case that the increased depth and more intricate architecture of ResNet adds unnecessary complexity to the extracted features when a more simple architecture (e.g., AlexNet) is sufficient for capturing environmental features relevant to species distributions. We note that object-centric image classification likely requires more discrimination than aerial image classification as there is more intra-class variability in object-centric images and infinitely-many image orientations as opposed to the fixed, bird's eye view of remote sensing images. To further explore the improved *AlexNet* features, we also extracted features from AlexNet models tuned to predict all classes present (*AlexNet_ind*) and the proportions of all present classes (*AlexNet_prop*) (Figure 4). In the Occupancy-Detection models, we observe slight (non-significant) improvements in the *AlexNet_ind* features and almost no difference between the *AlexNet_class* and *AlexNet_prop* features which indicates increasing the complexity of the training task is not always necessary.

Across all models, there is less difference between the deep features and the *eBird* features than we expected. Even more surprising is that the the *mean & SD* features perform as well as the *eBird* and deep features in Random Forest. However, given the similar performance of the *mean & SD* feature set and the *eBird* feature set, it stands to reason that there would be little further difference in the *eBird* and deep feature sets, as more complex habitat summaries do not appear to have much influence on predicting these species in Random Forest models. There are several potential explanations for these results. First, it is possible that simple habitat summaries are sufficient indicators of species occurrences in general. It is also possible that these generalist species in particular are less sensitive to habitat configuration and occur within a certain type of land cover regardless of the surroundings, in which case *mean & SD* would be sufficient and more complex feature sets may not contribute additional predictive power. Another possibility is that the chosen scale is not appropriate for determining how habitat configuration impacts the given species.

To investigate these potential hypotheses, future work will include analyzing specialist species known to be sensitive to habitat configuration to determine if the added complexity of the deep features is beneficial for predicting their distributions. Testing different scales is another direction of future work; this will require downloading entirely new datasets (at the new scales) and retraining the deep networks. A simple first pass could involve extracting deep features from multiple 2k × 2km patches surrounding the bird observation and averaging them to summarize larger habitats.

7 CONCLUSION

The preliminary results presented above do not show an advantage of deep features for species distribution modeling as clearly as we expected. The deep features showed quite similar performance in Random Forest models and more variability in the Occupancy-Detection models. While the deep environmental features do not consistently outperform simpler summary methods, on average, they perform as well as habitat summaries currently used in SDMs. However, given the power of deep learning in computer vision tasks, there is still reason to believe that features extracted by deep networks have the potential to outperform simpler habitat summaries in predicting species occurrences. We expect that our ongoing work to understand and extend these results (e.g., modeling species more sensitive to habitat type and configuration) will clarify best practices for extracting deep features from aerial images for predicting species distributions.

REFERENCES

[1] Matthew G. Betts, A. W. Diamond, G.J. Forbes, M.-A. Villard, and J.S. Gunn. 2006. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. *Ecological Modeling* 191 (January 2006), 197–224.
[2] Matthew G. Betts, Lenore Fahrig, Adam S. Hadley, Katherine E. Halstead, Jeff Bowman, W. Douglas Robinson, John A. Wiens, and David B. Lindenmayer. 2014. A species-centered approach for uncovering generalities in organism responses to habitat loss and fragmentation. *Ecography* 37, 6 (June 2014), 517–527.
[3] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (October 2001), 5–32.
[4] Wolfgang Buermann, Sassan Saatchi, Thomas B. Smith, Brian R. Zutta, Jaime A. Chaves, Borja Milá, and Catherine H. Graham. 2008. Predicting species distributions across the Amazonian and Andean regions using remote sensing data. *Journal of Biogeography* 35, 7 (July 2008), 1160–1176.
[5] Robert J. Jr Fletcher, Raphael K.Didham, Cristina Banks-Leite, Jos Barlow, Robert M. Ewers, James Rosindell, Robert D.Holt, Andrew Gonzalez, Renata Pardini, Ellen I. Damschen, Felipe P.L.Melo, Leslie Ries, Jayme A. Prevedello amd Teja Tscharntke, William F. Laurance, Thomas Lovejoy, and Nick M.Haddad. 2018. Is habitat fragmentation good for biodiversity? *Biological Conservation* 226 (October 2018), 9–15.

[6] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* (2017). https://doi.org/10.1016/j.rse.2017.06.031

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Las Vegas, Nevada, 770–778.

[8] Collin Homer, Jon Dewitz, Limin Yang, Suming Jin, Patrick Danielson, George Xian, John Coulston, Nathaniel Herold, James Wickham, and Kevin Megown. 2015. Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing* 81, 5 (May 2015), 345–354.

[9] Rebecca A. Hutchinson, Jonathon J. Valente, Sarah C. Emerson, Matthew G. Betts, and Thomas F. Dietterich. 2015. Penalized likelihood methods improve parameter estimates in occupancy models. *Methods in Ecology and Evolution* 6, 8 (August 2015), 949–959.

[10] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. 2019. Tile2Vec: unsupervised representation learning for spatially distributed data. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press, Honolulu, Hawaii.

[11] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 1725–1732.

[12] Jeremy T. Kerr and Marsha Ostrovsky. 2003. From space to species: ecological applications for remote sensing. *Trends in Ecology and Evolution* 18, 6 (June 2003), 299–305.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105.

[14] Darryl I. MacKenzie, James D. Nichols, Gideon B. Lachman, Sam Droege, and J. Andrew Royle. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83, 8 (2002), 2248–2255.

[15] K.S. McGarigal, Samuel Cushman, and Maile Neel E. Ene. 2012. FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. http://www.umass.edu/landeco/research/fragstats/fragstats.html.

[16] NAIP Information Sheet. 2015. https://www.fsa.usda.gov/Internet/FSA_File/naip_info_sheet_2015.pdf. Accessed: 2018-11-08.

[17] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. *CoRR* abs/1403.6382 (2014), 512–519.

[18] S. M. Shirley, Z. Yang, R. A. Hutchinson, J.D. Alexander, K. McGarigal, and M. G. Betts. 2013. Species distribution modelling for the people: unclassified landsat TM imagery predicts bird occurrence at fine resolutions. *Diversity and Distributions* 19, 7 (July 2013), 855–866.

[19] Brian L. Sullivan, Jocelyn L. Aycrigg, Jessie Barry, Rick E. Bonney, Nicholas E. Bruns, André A Dhondt, Andrew Farnsworth, John W Fitzpatrick, Thomas Fredericks, Jeff Gerbracht, Carla Gomes, Marshall J. Iliff, Carl Lagoze, Frank A. La Sorte, Matthew Merrifield, Mark Reynolds, Amanda D. Rodewald, Kenneth V. Rosenberg, Nancy M. Trautmann, David W. Winkler, Weng-Keen Wong, Jun Liang Yu, and Steve Kelling. 2014. The eBird enterprise : An integrated approach to development and application. https://doi.org/10.1016/j.biocon.2013.11.003

[20] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. 2016. Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, Phoenix, Arizona, 3929–3935.

[21] Jiaxuan You, Xiaocheng Li, Melvin Low, David B. Lobell, and Stefano Ermon. 2017. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press, San Francisco, California, 4559–4565.